STATISTICS IN MEDICINE
Statist. Med. 2009; **28**:2785–2800
Published online 9 July 2009 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.3660

The correlated and shared gamma frailty model for bivariate current status data: An illustration for cross-sectional serological data

N. Hens[1,2,*,†], A. Wienke[3], M. Aerts[1] and G. Molenberghs[1,4]

[1]*Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Diepenbeek, Limburg, Belgium*
[2]*Centre for Health Economics Research and Modeling Infectious Diseases, Centre for the Evaluation of Vaccination (WHO Collaborating Centre), Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium*
[3]*Institute of Medical Epidemiology, Biostatistics and Informatics, University Halle-Wittenberg, Halle, Germany*
[4]*Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Catholic University of Leuven, Leuven, Belgium*

SUMMARY

Frailty models are often used to study the individual heterogeneity in multivariate survival analysis. Whereas the shared frailty model is widely applied, the correlated frailty model has gained attention because it elevates the restriction of unobserved factors to act similar within clusters. Estimating frailty models is not straightforward due to various types of censoring. In this paper, we study the behavior of the bivariate-correlated gamma frailty model for type I interval-censored data, better known as current status data. We show that applying a shared rather than a correlated frailty model to cross-sectionally collected serological data on hepatitis A and B leads to biased estimates for the baseline hazard and variance parameters. Copyright © 2009 John Wiley & Sons, Ltd.

KEY WORDS: heterogeneity; correlation; current status; infectious diseases; bivariate binary data

1. INTRODUCTION

Analyzing time-to-event (TTE) data is not straightforward due to censoring. A particular type of censoring is interval censoring where event times are only known to lie in a specific interval. This situation especially happens when study subjects are not under continuous observation, for

*Correspondence to: N. Hens, Interuniversity Institute for Biostatistics and Statistical Bioinformatics, Hasselt University, Agoralaan 1, 3590 Diepenbeek, Belgium.
†E-mail: niel.hens@uhasselt.be

Contract/grant sponsor: SIMID; contract/grant number: 060081
Contract/grant sponsor: Belgian Government; contract/grant number: P6/03
Contract/grant sponsor: European Commission; contract/grant number: SSP22-CT-2004-502084
Contract/grant sponsor: German Research Council; contract/grant number: WI 3288/1-1

Received 11 December 2008
Accepted 8 June 2009

Copyright © 2009 John Wiley & Sons, Ltd.

example, patients visiting their doctor at predetermined times (or times that are convenient to them), where the occurrence of the event can be diagnosed knowing that the event has not occurred at the time of the last visit. Another situation is inspection times of technical equipment, where events can happen in between two inspection times. Consequently, it is only known that the event occurred between two visits or inspections, but not the exact time point. This kind of censoring is called interval censoring and was considered in detail by Sun [1] without special emphasis on frailty models. In general, right censoring (RC) is a special case of interval censoring and some of the methods for right-censored data can be directly, or with minor changes, applied to interval-censored data. However, most of the approaches for right-censored data are not appropriate for interval-censored data because the censoring mechanism behind interval censoring is much more complicated than in the case of right-censored data.

In this paper, we focus on case I interval-censored data better known as current status (CS) data, a term originating from demographical applications. That means that the observation for each individual survival time interval includes either zero or infinity. Such kind of data occur when each study subject is observed only once and the only available information for the event under study is whether the event has occurred before the observation was taken or not. Consequently, CS data are given in the form (T, Δ), where T denotes the inspection time and Δ is the indicator whether the event already occurred before the inspection or not.

There are many open research questions for the analysis of multivariate TTE data, which are much more challenging than their univariate counterparts. Available multivariate survival models fall into two broad classes—marginal and frailty models [2]. Marginal methods of analysis specify models for the effect of covariates on the hazards of the individual events (the margins), taking into account the fact that the observed event times are correlated but without the need for explicitly modeling this correlation [3]. The marginal approach is ideal for making inferences about the population average effect of risk factors on failure time. However, it provides limited insight into the multivariate relationship among failure times. These type of questions are answered by frailty models, explicitly considering the association between various events. In general, frailty models have an intuitive appeal and provide insight into the relationship between failures, and in this paper, we will zoom in on this approach.

A commonly used and very general approach to the problem of modeling multivariate data is to specify independence among observed data items conditional on a set of unobserved or latent variables (random effects). A multivariate model for the observed data is then induced by averaging over an assumed distribution for the latent variables. The dependence structure in the multivariate model arises when common or dependent latent variables enter into the conditional models for multiple observed data items. Frailty models for multivariate survival data are derived under a conditional independence assumption by specifying latent variables that act multiplicatively on the baseline hazard. This concept provides an extension of the traditional univariate frailty model [4, 5], and it allows to take the mutual dependence of life times of related individuals into account in the analysis of survival data.

There are two important approaches in this field, the shared frailty model and the correlated frailty model. In a shared frailty model, the frailty is common to the individuals in the group, and is thus responsible for creating dependence. The shared frailty model abounds in the literature on frailty models and was extensively studied in the monographs by Hougaard [6], Therneau and Grambsch [7] and Duchateau and Janssen [8].

The correlated frailty model is a natural extension of the shared frailty model. In the correlated frailty model, the frailties of individuals in a cluster are correlated, but not shared. It enables the

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

explicit inclusion of additional correlation parameters, whereas in the shared frailty approach all correlations between group members are equal.

In the following we will restrict our considerations to the bivariate case, because our motivating example is of bivariate nature. Extensions to higher-dimensional models are straightforward only in the shared frailty approach.

Various important research questions emerge when considering bivariate CS data. The first one deals with the question as to how much information is lost when CS data are observed instead of (right censored) life times. If the correlated frailty model is the underlying correct model, the obvious question becomes what is measured using the shared frailty approach on the CS data. Whereas the shared gamma frailty model was already applied to the CS data [9], very little is known about the correlated gamma frailty model in case of CS data [10, 11].

To answer these questions, simulations are performed based on the bivariate shared and correlated gamma frailty model under different censoring assumptions. We relate these results to the situation of infectious disease epidemiology where estimating the frailty variance based on cross-sectional serological data for multiple, similarly transmitted, pathogens is important to control the spread of infectious diseases [9].

We start by introducing a motivating example of multisera data on hepatitis A and B in Section 2. In Section 3, we introduce the correlated gamma frailty and its CS version. Furthermore, we describe how these CS frailty models are casted in the generalized linear mixed model framework. We fit these models to the hepatitis A and B data in Section 4. In Section 5, we examine the performance of the correlated gamma frailty for different types of censoring and the effect of ignoring the underlying correlation structure by using a shared rather than a correlated gamma frailty distribution. We end the paper with a discussion on the implications of modeling multivariate CS data using correlated frailty models and introduce topics for further research.

2. MOTIVATING EXAMPLE

Modeling infectious diseases is mostly done using compartmental models that describe the flow of individuals through different disease stages. One of the most important parameters in such a compartmental model describes the per capita rate at which a susceptible person acquires the infection and thus moves from the compartment of susceptible to the compartment of infected individuals. This per capita rate is the infection hazard and can be estimated from data on time to infection or incidence data. Collecting time to infection data or incidence data is hard and often unfeasible, because underreporting is likely to occur and follow-up studies are expensive and time consuming. Under the steady-state assumption and assuming lifelong immunity once infected, one can estimate the hazard of infection from cross-sectionally collected serological data. Serological data provide information on past infection and together with the individual's age (mostly registered in years) constitute CS data. The role of the TTE is assumed by the individual's age, given that a blood sample is taken at a specific point in time, and the time to infection, i.e. the time between birth and infection time, is actually the individual's age at the time of infection. These data are thus Type I censored.

The hazard of infection is often called the force of infection and can be considered as a reflection of the degree of contacts with transmission potential for the infection at hand. Often, data are from serological samples that are tested for more than one antigen. Such bivariate data make it possible to study the association between the acquisition of both infections [9, 12].

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

In the epidemic theory, Coutinho *et al.* [13] were the first to systematically treat heterogeneity in the acquisition of infections. Individuals are dissimilar in the way they acquire infections. Some individuals are more susceptible than others and will experience infection earlier. These frailties can be partly explained, but in most cases constitute an 'unexplained residual' component. Gaining insight in the frailty to acquire an infection has a potentially large impact on the design and implementation of control strategies.

Viral hepatitis is a serious health problem throughout the world. To obtain a clear picture of the prevalence of hepatitis A, B and C, a sero-epidemiological study was undertaken in 1993–1994 in Flanders. From the 4058 blood samples drawn in Flanders from a study group representative for the Flemish population, we focus on the complete cases and more specifically hepatitis A and B, resulting in 3787 blood samples. These blood samples were then tested for the presence of antibodies for the different infections and, using a pre-specified cut-off value, samples were classified as either positive or negative. Together with the patient's age and under the assumption of lifelong immunity, these data constitute CS data on whether or not past infection took place. Hepatitis C was not considered here because of its low prevalence (less than 1 per cent). More detailed information on these data can be found in Beutels *et al.* [14].

Next to age-dependent seroprofiles, it is of interest to look at the heterogeneity in acquisition of either infection and the correlation between the acquisition of both infections. While age-dependent seroprofiles reflect the age-specific risk of infection, the proper assessment of heterogeneity has direct implications with respect to the estimation of the basic reproduction number and the associated critical vaccination coverage [9]. Estimating the correlation in its own right could indicate transmission through similar routes (perfect correlation) or could reflect to what extent a latent process, such as the social or hygienic behavior of people, drives the more general infection process. Note that the main transmission route for hepatitis A is foodborne or faeco-oral and for hepatitis B is sexual or bloodborne, reflecting hygienic behavioral conduct of individuals. Moreover, co-infections, i.e. joint infections caused by more than one pathogen, are an aggravating factor in disease progression for virtually all infections and thus of interest to be quantified.

3. METHODS

In this section, we first introduce the shared frailty model as used by Farrington *et al.* [9] to model the heterogeneity in the acquisition of rubella and mumps in the U.K. We then propose the use of the correlated frailty model as an extension of the shared frailty model for the analysis of bivariate CS data.

Denote by $\lambda_i(t, Z_i)$ the hazard function at time t conditional on the frailty Z_i $(i=1,2)$. The corresponding conditional survival function $S_i(t|Z_i)$ $(i=1,2)$ is then given by

$$S_i(t|Z_i) = \mathrm{e}^{-\int_0^t \lambda_i(s,Z_i)\,\mathrm{d}s} \tag{1}$$

which we combine with the proportional hazards assumption $\lambda_i(t, Z_i) = Z_i \lambda_{i0}(t)$ to obtain

$$S_i(t|Z_i) = \mathrm{e}^{-Z_i \int_0^t \lambda_{i0}(s)\,\mathrm{d}s} \tag{2}$$

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

The unconditional survival function can be obtained by integrating out the random frailty Z_i by using the Laplace transform $\mathbf{L}_i$ of Z_i $(i=1,2)$:

$$S_i(t)=\mathbf{E}S_i(t|Z_i)=\mathbf{L}_i\left(\int_0^t \lambda_{i0}(s)\,\mathrm{d}s\right) \tag{3}$$

Assuming conditional independence, we can formulate the conditional bivariate survival function. Depending on the choice for the bivariate frailty distribution, either an explicit expression can be given or numerical integration is required. In general, numerical integration with respect to the frailty, or random-effects, distribution is not straightforward but has become more accessible through the development of appropriate statistical software and reformulating non-normal random effects, as done by Nelson *et al.* [15] and Liu and Yu [16]. In the following sections, we will focus on the gamma frailty distribution as the most often used frailty distribution because of its explicit solution for the unconditional survival function (see e.g. [6, 8]), which owes to conjugacy properties.

3.1. The shared gamma frailty model

In the shared gamma frailty model, the bivariate frailty distribution (Z_1, Z_2) is characterized by $Z=Z_1=Z_2$. The unconditional bivariate survival function is given by

$$S(t_1,t_2)=[S_1^{-\sigma^2}(t_1)+S_2^{-\sigma^2}(t_2)-1]^{-1/\sigma^2} \tag{4}$$

Here σ^2 represents the variance of Z.

3.2. The correlated gamma frailty model

Although the shared gamma frailty model assumes perfect correlation and a common variance, the correlated gamma frailty model as introduced in Yashin *et al.* [17] is more flexible. These authors used an additive decomposition of the frailty variables into the sum of independent gamma distributed variables to construct a bivariate frailty distribution.

The bivariate frailty distribution may be constructed using independent additive components $Y_i, i=0,1,2$ with one component common to both frailties (i.e. $Z_i=\sigma_i^2(Y_0+Y_i)$, $i=1,2$), introducing an additional parameter characterizing the correlation between the frailties [17], hence the name 'correlated frailty models', and by multiplying with σ_i^2, $(i=1,2)$ restricting the mean to one while allowing for different variances. More specifically, assuming that k_0, k_1 and k_2 are some real-positive parameters, $Y_i\sim\Gamma(k_i,1)\,(i=0,1,2)$ and $\sigma_i^2=(k_0+k_i)^{-1}$ $(i=1,2)$ and $\rho=k_0[(k_0+k_1)(k_0+k_2)]^{-1/2}$. Note that $k_i\geqslant 0\,(i=0,1,2)$ implies $\sigma_i^2>0$ $(i=1,2)$ and $0\leqslant\rho\leqslant\min(\sigma_1\sigma_2^{-1},\sigma_2\sigma_1^{-1})$. The identifiability of the correlated frailty model for bivariate event times without covariates has been established before [18].

The explicit expression for the survival function in terms of σ_1, σ_2 and ρ is given by Yashin *et al.* [17]:

$$S(t_1,t_2)=[S_1(t_1)]^{1-(\sigma_1/\sigma_2)\rho}[S_2(t_2)]^{1-(\sigma_2/\sigma_1)\rho}[S_1^{-\sigma_1^2}(t_1)+S_2^{-\sigma_2^2}(t_2)-1]^{-\rho/\sigma_1\sigma_2} \tag{5}$$

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Note that if $Z_1 = Z_2$, and thus $\sigma_1 = \sigma_2 = \sigma, \rho = 1$, we end up with the shared gamma frailty model (4).

3.3. Current status data

Before turning to the CS likelihood function, let us write down the likelihood function for both the uncensored and right-censored TTE situation. The likelihood function for uncensored TTE data is given by

$$L(t_1, t_2) = \frac{\partial^2}{\partial t_1 \partial t_2} S(t_1, t_2) \tag{6}$$

where $S(t_1, t_2)$ is given by (5). Although this likelihood function is relatively easy to derive, we are omitting the somewhat awkward expression here.

There exist different types of censoring of which the most common is RC. Let us define the censoring indicator δ_{ij}, which takes value 1 if the individual j has experienced the event i, and 0 otherwise. The corresponding likelihood function can then be derived by:

$$L(t_1, t_2, \delta_1, \delta_2) = \delta_1 \delta_2 \left[\frac{\partial^2}{\partial t_1 \partial t_2} S(t_1, t_2) \right] + \delta_1 (1 - \delta_2) \left[-\frac{\partial}{\partial t_1} S(t_1, t_2) \right]$$

$$+ (1 - \delta_1) \delta_2 \left[-\frac{\partial}{\partial t_2} S(t_1, t_2) \right] + (1 - \delta_1)(1 - \delta_2) S(t_1, t_2) \tag{7}$$

In the case of bivariate CS data, the likelihood function can easily be expressed in terms of the unconditional bivariate and univariate survival functions [1]:

$$L(t_1, t_2, \delta_1, \delta_2) = \delta_1 \delta_2 [1 - S_1(t_1) - S_2(t_2) + S(t_1, t_2)] + \delta_1 (1 - \delta_2) [S_2(t_2) - S(t_1, t_2)]$$

$$+ (1 - \delta_1) \delta_2 [S_1(t_1) - S(t_1, t_2)] + (1 - \delta_1)(1 - \delta_2) S(t_1, t_2) \tag{8}$$

with $S_1(t_1) = S(t_1, 0)$ and $S_2(t_2) = S(0, t_2)$ as marginal survival functions. We will use function (5) as bivariate survival function in the following. Note that in case of CS data without any covariates, the model is not identifiable using a nonparametric baseline hazard [11], motivating the use of a parametric baseline hazard function such as, for example, the Gompertz baseline hazard where $\lambda_{i0}(t) = a_i \exp(b_i t), i = 1, 2$. Although we do not investigate the sufficient conditions required for the model to be identifiability in this paper, we rely on the more general methodology of detecting parameter redundancy [19, 20].

Note that, when assuming univariate monitoring times $t = t_1 = t_2$, the link to generalized linear mixed models is readily established. Indeed, looking at (2), the corresponding generalized linear mixed model is given by $(i = 1, 2)$: $\log[S_i(t)] = -Z_i \Lambda_i(t)$, where (Z_1, Z_2) is the bivariate random effect of which each component acts multiplicatively on the (cumulative) hazard function $\Lambda_i(t) = \int_0^t \lambda_{i0}(s)\, \mathrm{d}s$ for $i = 1, 2$. As in case of the correlated frailty model, different choices of the correlated random effects distribution can be made such as the mathematically convenient gamma distribution, which leads to an explicit expression of the unconditional multinomial likelihood.

The link between frailty models and generalized linear mixed models has been established before in various settings such as in the proportional hazards model for clustered survival data

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

(see e.g. [21–23]). For an extended overview of the generalized linear mixed model and its applications we refer to the literature [24–27].

3.4. An overall association measure

Measures of association are essential tools for the analysis of bivariate data. Among the most familiar is Kendall's tau. Betensky and Finkelstein [28] proposed an imputation-based estimator for Kendall's tau in the case of bivariate interval-censored data. This estimator imputes TTEs based on the estimated survivor function while assigning a zero score to overlapping rectangles, i.e. bivariate interval-censored data are often conceived as rectangular in the plane, because no ordering is possible whenever rectangles overlap. Note that for Type I interval-censored data the number of overlapping rectangles is abundant because of the inclusion of either zero or infinity in both dimensions. Because of sparse information this method is no longer applicable for the specific case of Type I interval-censored data with univariate monitoring times, i.e. the current status data that are the focus of this paper. Therefore, following the philosophy of generating data from an estimated survivor function, we define a simulation-based estimate of Kendall's tau for the various censoring schemes.

It is a straightforward idea to, based on the estimated model for any censoring scheme, generate new simulated data that can then be used to derive a simulation-based empirical estimate of Kendall's tau by looking at the concordance score of each pair $(i, j), i \neq j = 1, \ldots, n$ of bivariate observations $\{(T_{1i}, T_{1j}), (T_{2i}, T_{2j})\}$. Whenever $T_{1i} > T_{2i}$ and $T_{1j} > T_{2j}$ or $T_{1i} < T_{2i}$ and $T_{1j} < T_{2j}, (T_{1i}, T_{1j})$ and (T_{2i}, T_{2j}) are called concordant, whereas $T_{1i} > T_{2i}$ and $T_{1j} < T_{2j}$ or $T_{1i} < T_{2i}$ and $T_{1j} > T_{2j}, (T_{1i}, T_{1j})$ and (T_{2i}, T_{2j}) are called disconcordant. Based on these definitions, Kendall's tau can be calculated as the difference of the probability of concordance and discordance:

$$\tau = P\{(T_{1i} - T_{1j})(T_{2i} - T_{2j}) > 0\} - P\{(T_{1i} - T_{1j})(T_{2i} - T_{2j}) < 0\} \tag{9}$$

As a result, the estimate of τ will be made over several simulations, thence the simulation-based standard error can be calculated as well. Note that calculating (9) can be done explicitly, but was found practically prohibitive given that no closed-form solution exists and numerical integration proved to be tedious.

4. APPLICATION TO MULTISERA DATA ON HEPATITIS A AND B

Table I summarizes the results of applying the correlated gamma frailty model, the correlated gamma frailty model with equal variances, the shared gamma frailty model and a model without frailty, assuming independence and no heterogeneity, to the multisera data on hepatitis A and B introduced in Section 2. The corresponding SAS-code can be found in the Appendix. While the loglikelihood function favors the unrestricted correlated frailty model, the observed difference with the correlated frailty model assuming equal variances is non-significant based on the corresponding likelihood ratio test ($p = 0.655$). When comparing the correlated frailty model with equal variances to the shared frailty model, we are interested in testing the null hypothesis H_0: $\rho = 1$ vs H_1: $\rho < 1$, which lies on the boundary of the parameter space. Therefore, the limiting distribution follows a 50:50 mixture of a χ^2_0 and χ^2_1 distribution [29, 30]. The corresponding $p < 0.001$ clearly favors the alternative hypothesis. A comparison of the correlated

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table I. Parameter estimates and standard errors for the Hepatitis A and B analyses result using a Gompertz baseline hazard ($\lambda_{i0}(t)=a_i \exp(b_i t), i=1,2$) and various versions of the correlated gamma frailty model.

Restrictions	Unrestricted	Equal variances $\sigma_1=\sigma_2$	Shared frailty $\sigma_1=\sigma_2,\ \rho=1$	Univariate frailty $\rho=0$	Independence $\sigma_1=\sigma_2=0$
a_1	0.007 (0.001)	0.007 (0.001)	0.012 (0.001)	0.008 (0.001)	0.015 (0.001)
b_1	0.105 (0.017)	0.104 (0.017)	0.037 (0.005)	0.086 (0.016)	0.019 (0.019)
a_2	0.002 (4E−4)	0.002 (4E−4)	0.002 (3E−4)	0.002 (3E−4)	0.002 (3E−4)
b_2	0.000 (0.008)	0.002 (0.008)	−0.000 (0.007)	0.000 (0.007)	−0.002 (0.007)
σ_1	1.632 (0.501)	1.628 (0.174)	0.723 (0.084)	1.422 (0.180)	0.000 (–)
σ_2	1.167 (0.174)	1.628 (0.174)	0.723 (0.084)	1.016 (2.557)	0.000 (–)
ρ	0.677 (0.283)	0.487 (0.079)	1.000 (–)	0.000 (–)	0.000 (–)
-2ℓ	5653.4	5653.6	5687.0	5682.9	5713.1

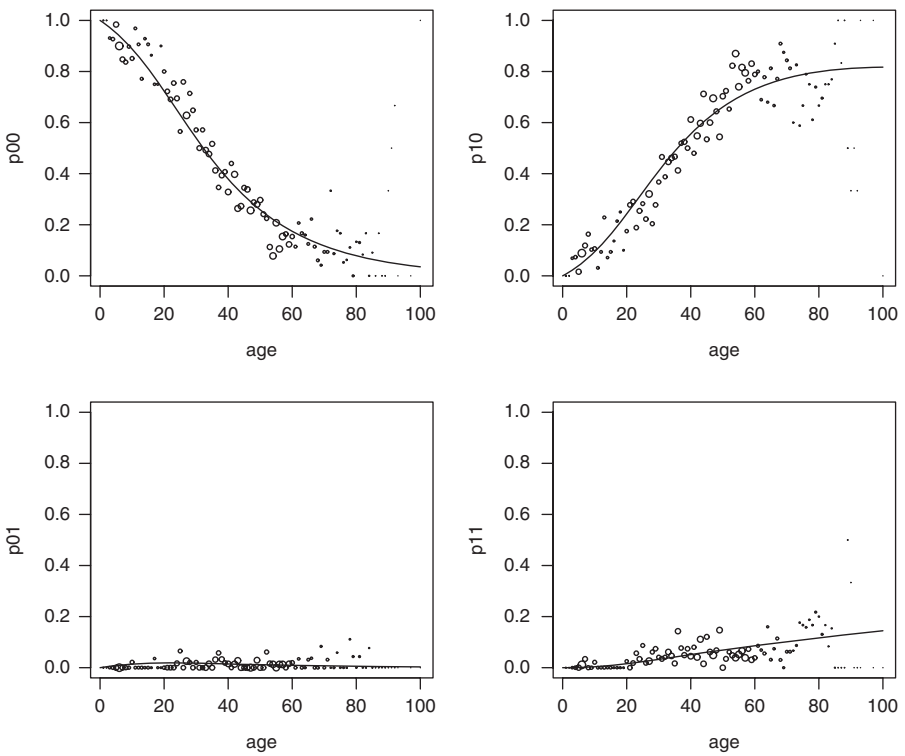

Figure 1. Plot of the joint probabilities of hepatitis A and B and the correlated frailty fit with equal variances. $p00$ refers to the joint probability of no past infection for either virus (left upper panel); $p10$ refers to past and no past infection for hepatitis A and B, respectively (right upper panel); $p01$ refers to no past and past infection for hepatitis A and B, respectively; and $p11$ refers to past infection for both viruses.

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

frailty with common variance with the result of a univariate analysis assuming $\rho=0$ results in a similar conclusion ($p<0.001$). Figure 1 shows the estimated and observed joint probabilities based on the correlated frailty model with equal variances, visually indicating a good fit to the data.

5. SIMULATIONS

To gain more insight in how much information is lost when turning from TTE data to RC and finally to CS data, we first performed a simulation study. Second, we used a simulation study to investigate the impact of misspecifying the frailty distribution. We started by generating bivariate TTE data based on the bivariate gamma frailty model and Gompertz baseline hazards. Generating the correlated gamma frailty (Z_1, Z_2) is done via its additive components $Y_i, i=0,1,2$. Given Z_i, the TTE T_i is generated by calculating $\log[1-b_i \log(u_i)/(a_i z_i)]/b_i$ where u_i is generated from the uniform distribution U(0, 1) and a_i, b_i are the parameters from the Gompertz baseline hazard $\lambda_{i0}(t)=a_i \exp(b_i t)$ $(i=1,2)$. The censoring indicator Δ_i $(i=1,2)$ was then generated by comparing the censoring time $T_i^{\Delta} \sim \mathrm{U}(0,75)$ $(i=1,2)$ to the generated TTE T_i $(i=1,2)$. Whenever $T_i^{\Delta}>T_i$, $\Delta_i=1$ and 0 otherwise $(i=1,2)$. Right-censored data were obtained by $\{\min(T_i, T_i^{\Delta}), \Delta_i\}$ $(i=1,2)$, whereas $\{T_i^{\Delta}, \Delta_i\}$ $(i=1,2)$ constituted Type I interval-censored data.

The choice of the Gompertz baseline with parameters a_i, b_i $(i=1,2)$ and the sample size of 3787 was inspired by the hepatitis A and B example, multisera data that typically have rather a large sample size and for which the baseline hazard can plausibly be assumed to be of the Gompertz type. Five hundred data sets were generated and analyzed for all settings. Note that other simulations with smaller heterogeneity parameters showed similar results and are available in the Appendix.

5.1. Censoring

In a first part, the simulation study aims at identifying the information loss when transferring TTE to RC and CS. Table II shows the true values, the parameter estimates and the empirical standard errors (e.s.e.) for the different censoring schemes. It is observed that empirical standard errors increase with increasing information loss (TTE $\rightarrow$ RC $\rightarrow$ CS), whereas the estimates show consistency when increasing the sample size toward 10 000 observations and more (see Appendix Tables AI and AII). Parameters are chosen so that around 21 per cent of the observations are censored (censored data) or state that the infection not occurred yet (CS data).

Estimating Kendall's τ using the simulation procedure as outlined in Section 3.4 resulted in $\hat{\tau}_{\mathrm{tte}}=0.169$ (e.s.e. 0.016) for the uncensored TTE setting; $\hat{\tau}_{\mathrm{rc}}=0.173$ (e.s.e. 0.034) for the RC situation and $\hat{\tau}_{\mathrm{cs}}=0.181$ (e.s.e. 0.052) for the CS data. Again, the estimated values $\hat{\tau}_{\mathrm{tte}}, \hat{\tau}_{\mathrm{rc}}$ and $\hat{\tau}_{\mathrm{cs}}$ do not differ substantially, while the corresponding standard error not surprisingly increases when information is lost.

5.2. Misspecification

We investigate the effect when misspecifying the assumed frailty distribution using the CS data as generated above while analyzing the data assuming a correlated frailty distribution with common

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table II. Averaged parameter estimates and empirical standard errors for the simulation study of the correlated gamma frailty model with uncensored time to event; right-censored and current status data using a Gompertz baseline hazard ($\lambda_{i0}(t)=a_i \exp(b_i t), i=1,2$).

Parameter	True value	Uncensored time to event mean (e.s.e.)	Right-censored data mean (e.s.e.)	Current status data mean (e.s.e.)
a_1	0.006	0.006 (0.001)	0.006 (0.001)	0.006 (0.001)
b_1	0.020	0.020 (0.002)	0.022 (0.010)	0.045 (0.420)
a_2	0.008	0.008 (0.001)	0.008 (0.001)	0.008 (0.001)
b_2	0.030	0.030 (0.003)	0.032 (0.007)	0.048 (0.228)
σ_1	1.600	1.604 (0.113)	1.621 (0.466)	1.694 (1.854)
σ_2	1.000	0.999 (0.068)	1.056 (0.214)	1.179 (0.920)
ρ	0.500	0.501 (0.035)	0.540 (0.169)	0.636 (0.257)

Table III. Averaged parameter estimates and empirical standard errors for the simulation study on the misspecification of the frailty distribution for current status data using a Gompertz baseline hazard ($\lambda_{i0}(t)=a_i \exp(b_i t), i=1,2$).

Parameter	True value	Correlated frailty mean (e.s.e.)	Common variance CF mean (e.s.e.)	Shared frailty mean (e.s.e.)	Univariate frailty mean (e.s.e.)
a_1	0.006	0.006 (0.001)	0.006 (0.001)	0.006 (0.001)	0.006 (0.005)
b_1	0.020	0.045 (0.420)	0.013 (0.009)	0.007 (0.004)	0.062 (0.135)
a_2	0.008	0.008 (0.001)	0.008 (0.001)	0.008 (0.001)	0.008 (0.001)
b_2	0.030	0.048 (0.228)	0.039 (0.019)	0.024 (0.003)	0.047 (0.047)
σ_1	1.600	1.694 (1.854)	1.185 (0.429)	0.769 (0.051)	1.962 (2.219)
σ_2	1.000	1.179 (0.920)	1.185 (0.429)	0.769 (0.051)	1.107 (0.941)
ρ	0.500	0.636 (0.257)	0.679 (0.219)	1.000 (–)	0.000 (–)

variances, a shared frailty distribution and two univariate frailty distributions. For completeness, we also repeat the results of the correlated frailty distribution analysis while summarizing the parameter estimates and empirical standard errors in Table III.

Although the estimates of the Gompertz baseline parameters can be considered stable over the different models, there is quite some difference between the estimated variance parameters. This is not surprising, given what Wienke *et al.* [31] observed before, i.e. the negative correlation between ρ and σ_i $(i=1,2)$. Indeed, the correlated frailty with common variance estimates the variance parameters to be smaller at the cost of a larger correlation when compared with the correlated frailty model. The shared frailty, assuming perfect correlation and common variance, results in a lower variance estimate when compared with the common variance correlated frailty model. Finally, the univariate frailties, assuming independence and thus zero correlation, result in a substantially higher and comparable variance estimate.

Again, using simulations and the procedure outlined in Section 3.4, Kendall's τ was estimated as 0.225 (e.s.e. 0.036) for the model with shared frailty; and 0.205 (e.s.e. 0.034) for the model with common variance correlated frailty, both of which exceed the correlated frailty-based estimate.

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table IV. Averaged parameter estimates and empirical standard errors for the simulation study of the correlated gamma frailty model with uncensored time to event; right-censored and current status data using a Gompertz baseline hazard ($\lambda_{i0}(t) = a_i \exp(b_i t), i = 1, 2$).

Parameter	True value	Uncensored time to event mean (e.s.e.)	Right-censored data mean (e.s.e.)	Current status data
a_1	0.006	0.006 (2E−4)	0.006 (3E−4)	0.006 (0.001)
b_1	0.020	0.020 (0.001)	0.022 (0.007)	0.025 (0.018)
a_2	0.008	0.008 (3E−4)	0.008 (4E−4)	0.008 (0.001)
b_2	0.030	0.030 (0.002)	0.032 (0.004)	0.034 (0.009)
σ_1	0.800	0.799 (0.048)	0.852 (0.339)	0.897 (0.575)
σ_2	0.600	0.599 (0.040)	0.642 (0.164)	0.701 (0.280)
ρ	0.500	0.502 (0.052)	0.546 (0.241)	0.595 (0.292)

Table V. Averaged parameter estimates and empirical standard errors for the simulation study on the misspecification of the frailty distribution for current status data using a Gompertz baseline hazard ($\lambda_{i0}(t) = a_i \exp(b_i t), i = 1, 2$).

Parameter	True value	Correlated frailty mean (e.s.e.)	Common variance CF mean (e.s.e.)	Shared frailty mean (e.s.e.)	Univariate frailty mean (e.s.e.)
a_1	0.006	0.006 (0.001)	0.006 (0.001)	0.006 (0.001)	0.006 (0.001)
b_1	0.020	0.025 (0.018)	0.022 (0.015)	0.016 (0.003)	0.054 (0.428)
a_2	0.008	0.008 (0.001)	0.008 (0.001)	0.008 (0.001)	0.008 (0.001)
b_2	0.030	0.034 (0.009)	0.034 (0.009)	0.027 (0.002)	0.034 (0.012)
σ_1	0.800	0.897 (0.575)	0.770 (0.484)	0.453 (0.056)	1.049 (1.745)
σ_2	0.600	0.701 (0.280)	0.770 (0.484)	0.453 (0.056)	0.578 (0.458)
ρ	0.500	0.595 (0.292)	0.667 (0.310)	1.000 (–)	0.000 (–)

5.3. Simulation setting 2

In a second simulation setting, data were generated using similar Gompertz baseline parameters but with smaller heterogeneity parameters and a correlation of 0.50. Again, we investigate how much information is lost when turning from TTE to RC and finally to CS, and what the impact of misspecifying the frailty distribution is. Five hundred data sets of size 5000 were generated and analyzed. Tables IV and V show similar results when compared with those of the first simulation setting.

6. DISCUSSION

Analysis of multivariate survival time data provides an exciting example for challenging modeling strategies. Available statistical models fall into two broad classes—marginal and frailty models. Marginal methods consider the association between the events as a nuisance parameter. The other

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

commonly used and very general approach to multivariate survival data is to specify independence among observed data items conditional on a set of unobserved or latent variables (random effects) which act multiplicatively on the baseline hazard. Especially shared frailty models have a long tradition in modeling clustered survival times. However, shared frailty models have some limitations.

First, the concept of shared frailty forces the unobserved factors to be the same within the cluster, which is generally inappropriate when interested in disentangling association and heterogeneity. Indeed, in general, it may be inappropriate to assume that both infections considered in the example above share all of their unobserved risk factors. However, in practice we note that it is possible to combine the unobserved factors into one frailty, which is sufficiently rich in distribution.

Second, the dependence parameter and the population heterogeneity are confounded. Elbers and Ridder [32] showed that this problem exists for any univariate frailty distribution with a finite mean. However, 'shared frailty' in bivariate and multivariate models differs from 'individual frailty' used in the case of univariate data. Initially this difference in the notions of frailty was not clearly understood. It is worth noting that the value of σ^2 estimated from the univariate data may, in fact, have nothing to do with association.

To circumvent these problems, correlated frailty models were established to have different parameters for the association and heterogeneity. The present paper applies this approach to the problem of CS data with special focus on hepatitis A and B. Using the correlated gamma frailty model instead of the shared gamma frailty model leads to significant improvement of the likelihood, which speaks in favor for the former model. An additional advantage is the nice interpretation of the parameters. Here σ_1 and σ_2 are the measures of population heterogeneity in the susceptibility to hepatitis A and B, respectively. Furthermore, the parameter ρ—even if not the correlation between the original event times—can be interpreted as a correlation measure. These parameters yield implications for further programs to prevent the infections as the critical vaccination coverage is higher for more heterogenous populations [33].

There exists a strong link between copula models and frailty models. Besides the fact that frailty models and copulas look very similar, it is important to note that there are also differences between both approaches, which are often overlooked. In the application presented in this paper, we used a gamma frailty that directly relates to a Clayton copula and brings together frailty and copula models. However, copula models, often used to assess the association between event times, in general, cannot capture the heterogeneity as frailty models can. For more details regarding this aspect we refer to the paper by Goethals *et al.* [34]. Bivariate copula models for CS data including limiting distributions are discussed in Wang and Ding [35].

Further research is needed to investigate the impact of various misspecifications on the correlated frailty model. White's theory about inferences in miss-specified models can be employed to investigate such sources of miss-specification, using his likelihood-type or Lagrange-multiplier tests [36]. Note that this theory can also be regarded as the underpinning of such commonly used methods as generalized estimating equations. This methodology has been applied by Litiére *et al.* [37, 38] in the context of miss-specification arising from the random-effects distribution in generalized linear mixed models. Moreover, we did not address the effect of ignoring the correlated frailty structure by using a shared frailty distribution on the estimation of covariate effects. Similarly, further research is needed to investigate how copulas can be used to model bivariate CS data and on the extension toward multivariate CS data.

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

APPENDIX A

A.1. SAS-code to fit the correlated gamma frailty to current status data

SAS-code to fit the correlated gamma frailty to the bivariate data on past infection of hepatitis A and B. The code uses a reparametrization of the variance parameters in terms of the additive decomposition of the correlated gamma frailty. Data are organized in rectangular format with on each line the age of the individual and response 1: no previous infection for either disease, 2: previous infection for hepatitis A only, 3: previous infection for hepatitis B only and 4: previous infection for hepatitis A and B.

```
proc nlmixed data=hepdata tech=congra maxiter=5000 gconv=1e-22;
parms a1eta=-5 b1=0.1045 a2eta=-6 b2=-0.0022 k0eta=-2 k1eta=-2 k2eta=-2;
title 'Correlated frailty model to hepdata';
a1=exp(a1eta);
a2=exp(a2eta);
k0=exp(k0eta);
k1=exp(k1eta);
k2=exp(k2eta);
sigma1=1/sqrt(k0+k1);
sigma2=1/sqrt(k0+k2);
rho=k0/sqrt((k0+k1)*(k0+k2));
clambdaHAV=a1/b1*(exp(b1*age)-1);
clambdaHBV=a2/b2*(exp(b2*age)-1);
S1a=(1+sigma1**2*clambdaHAV)**(-1/sigma1**2);
S2a=(1+sigma2**2*clambdaHBV)**(-1/sigma2**2);
S12a=((S1a**(-sigma1**2)+S2a**(-sigma2**2)-1)**(-rho/(sigma1*sigma2)));
p00=S1a**(1-sigma1/sigma2*rho)*S2a**(1-sigma2/sigma1*rho)*S12a;
p10=S2a-p00;
p01=S1a-p00;
p11=1-p00-p10-p01;
ll=(response=1)*log(p00)+(response=2)*log(p10)+(response=3)*log(p01)+
(response=4)*log(p11);
model response general(ll);
estimate 'a1' exp(a1eta);
estimate 'b1' b1;
estimate 'a2' exp(a2eta);
estimate 'b2' b2;
estimate 'sigma1' 1/sqrt(k0+k1);
estimate 'sigma2' 1/sqrt(k0+k2);
estimate 'rho' k0/sqrt((k0+k1)*(k0+k2));
estimate 'k0' exp(k0eta);
estimate 'k1' exp(k1eta);
estimate 'k2' exp(k2eta);
run;
```

A.2. Extra simulations

In this section we document the results of the simulation study for sample size 10 000 (Tables AI and AII).

Copyright © 2009 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table AI. Averaged parameter estimates and empirical standard errors for the simulation study of the correlated gamma frailty model with uncensored time to event; right-censored and current status data using a Gompertz baseline hazard ($\lambda_{i0}(t) = a_i \exp(b_i t), i = 1, 2$) with sample size 10 000.

Parameter	True value	Uncensored time to event mean (e.s.e.)	Right-censored data mean (e.s.e.)	Current status data mean (e.s.e.)
a_1	0.006	0.006 (2E−4)	0.006 (2E−4)	0.006 (0.001)
b_1	0.020	0.020 (0.001)	0.021 (0.006)	0.022 (0.014)
a_2	0.008	0.008 (2E−4)	0.008 (3E−4)	0.008 (0.001)
b_2	0.030	0.030 (0.001)	0.031 (0.004)	0.032 (0.009)
σ_1	1.600	1.604 (0.054)	1.610 (0.290)	1.583 (0.589)
σ_2	1.000	1.002 (0.033)	1.020 (0.130)	1.047 (0.240)
ρ	0.500	0.501 (0.016)	0.516 (0.099)	0.581 (0.211)

Table AII. Averaged parameter estimates and empirical standard errors for the simulation study on the misspecification of the frailty distribution for current status data using a Gompertz baseline hazard ($\lambda_{i0}(t) = a_i \exp(b_i t), i = 1, 2$) with sample size 10 000.

Parameter	True value	Correlated frailty mean (e.s.e.)	Common variance CF mean (e.s.e.)	Shared frailty mean (e.s.e.)	Univariate frailty mean (e.s.e.)
a_1	0.006	0.006 (0.001)	0.006 (3E−4)	0.006 (3E−4)	0.006 (0.001)
b_1	0.020	0.022 (0.014)	0.014 (0.005)	0.007 (0.002)	0.008 (0.001)
a_2	0.008	0.008 (0.001)	0.008 (4E−4)	0.008 (3E−4)	0.032 (0.428)
b_2	0.030	0.032 (0.009)	0.033 (0.007)	0.024 (0.001)	0.029 (0.012)
σ_1	1.600	1.583 (0.589)	1.267 (0.263)	0.769 (0.022)	0.960 (1.829)
σ_2	1.000	1.047 (0.240)	1.267 (0.263)	0.769 (0.022)	0.840 (0.426)
ρ	0.500	0.581 (0.211)	0.705 (0.170)	1.000 (–)	0.000 (–)

ACKNOWLEDGEMENTS

We thank the associate editor and both reviewers for their comments leading to an improved presentation of the paper. We thank Philippe Beutels for making the data available to us and Tom Cattaert for the insightful discussions on the matter. This work has been funded by 'SIMID', a strategic basic research project funded by the institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), project number 060081 and by the IAP research network nr P6/03 of the Belgian Government (Belgian Science Policy). This work has benefitted from discussion held in POLYMOD, a European Commission project funded within the Sixth Framework Programme, contract number: SSP22-CT-2004-502084. Andreas Wienke was supported by the German Research Council, project number WI 3288/1-1.

REFERENCES

1. Sun J. *The Statistical Analysis of Interval-censored Failure Time Data*. Springer: New York, 2005.
2. Wei L, Glidden D. An overview of statistical methods for multiple failure time data in clinical trials. *Statistics in Medicine* 1997; **16**:831–839. DOI: 10.1002/(SICI)1097-0258(19970430)16:8<833::AID-SIM538>3.0.CO;2-2.
3. Wei L, Lin D, Weissfeld L. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association* 1989; **84**:1065–1073.
4. Vaupel J, Manton K, Stallard E. The impact of heterogeneity in individual frailty in the dynamics of mortality. *Demography* 1979; **16**(3):439–454.
5. Lancaster T. Econometric methods for the duration of unemployment. *Econometrica* 1979; **47**:939–956.
6. Hougaard P. *Analysis of Multivariate Survival Data*. Springer: New York, 2000.

Copyright © 2009 John Wiley & Sons, Ltd.

7. Therneau T, Grambsch P. *Modelling Survival Data*. Springer: Berlin, 2000.
8. Duchateau L, Janssen P. *The Frailty Model*. Springer: Berlin, 2008. DOI: 10.1007/978-0-387-72835-3.
9. Farrington C, Kanaan M, Gay N. Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. *Applied Statistics* 2001; **50**:251–292. DOI: 10.1111/1467-9876.00233.
10. Giard N, Lichtenstein P, Yashin A. A multistate model for the genetic analysis of the ageing process. *Statistics in Medicine* 2002; **21**:2511–2526. DOI: 10.1002/sim.1197.
11. Chang I, Wen C, Wu Y. A profile likelihood theory for the correlated gamma-frailty model with current status family data. *Statistica Sinica* 2007; **17**:1023–1046.
12. Hens N, Aerts M, Shkedy Z, Theeten H, Van Damme P, Beutels P. Modelling multi-sera data: the estimation of new joint and conditional epidemiological parameters. *Statistics in Medicine* 2008; **27**:2651–2664. DOI: 10.1002/sim.3089.
13. Coutinho F, Massad E, Lopez L, Burattini M, Struchiner C, Azevedo-Neto R. Modelling heterogeneities in individual frailties in epidemic models. *Mathematical and Computer Modelling* 1999; **30**:97–115.
14. Beutels M, Van Damme P, Aelvoet W, Desmyter J, Dondeyne F, Goilav C, Mak R, Muylle L, Pierard D, Stroobant A, Van Loock F, Waumans P, Vranckx R. Prevalence of hepatitis A, B and C in the flemish population. *European Journal of Epidemiology* 1997; **13**:275–280. DOI: 10.1023/A:1007393405966.
15. Nelson K, Lipsitz S, Fitzmaurice G, Ibrahim JG, Parzen M, Strawderman R. Use of the probability integral transformation to fit nonlinear mixed-effects models with nonnormal random effects. *Journal of Computational and Graphical Statistics* 2006; **15**:39–57. DOI: 10.1198/106186006X96854.
16. Liu L, Yu Z. A likelihood reformulation method in non-normal random effects models. *Statistics in Medicine* 2007; **27**:3105–3124. DOI: 10.1002/sim.3153.
17. Yashin A, Vaupel J, Iachine I. Correlated individual frailty: an advantageous approach to survival analysis of bivariate data. *Mathematical Population Studies* 1995; **5**:145–159.
18. Iachine IA, Yashin AI. Identifiability of bivariate frailty models based on additive independent components. *Research Report 8*, Department of Statistics and Demography, Odense University, 1998.
19. Catchpole E, Morgan B. Detecting parameter redundancy. *Biometrika* 1997; **84**:187–196. DOI: 10.1093/biomet/84.1.187.
20. Catchpole E, Morgan B. Deficiency of parameter redundant models. *Biometrika* 2001; **88**:593–598. DOI: 10.1093/biomet/88.2.593.
21. Vaida F, Xu R. Proportional hazards model with random effects. *Statistics in Medicine* 2000; **24**:3309–3324. DOI: 10.1002/1097-0258(20001230)19:24<3309::AID-SIM825>3.0.CO;2-9.
22. Yau KKW. Multilevel models for survival analysis with random effects. *Biometrics* 2004; **57**:96–102. DOI: 10.1111/j.0006-341X.2001.00096.x.
23. Yau KKW, McGilchrist C. Use of generalised linear mixed models for the analysis of clustered survival data. *Biometrical Journal* 2007; **39**:1–11. DOI: 10.1002/bimj.4710390102.
24. Molenberghs G, Verbeke G. *Models for Discrete Longitudinal Data*. Springer: New York, 2005.
25. Jiang J. *Linear and Generalized Linear Mixed Models and their Applications*. Springer Series in Statistics. Springer: New York, 2007.
26. Massonnet G, Janssen P, Burzykowski T. Fitting conditional survival models to meta-analytic data by using a transformation toward mixed-effects models. *Biometrics* 2008; **64**(3):834–842. DOI: 10.1111/j.1541-0420.2007.00960.x.
27. McCulloch C, Searle S, Neuhaus J. *Generalized, Linear, and Mixed Models*. Wiley: New York, 2008.
28. Betensky R, Finkelstein D. An extension of Kendall's coefficient of concordance to bivariate interval censored data. *Statistics in Medicine* 1999; **18**:3101–3109. DOI: 10.1002/(SICI)1097-0258(19991130)18:22<3101::AID-SIM339>3.0.CO;2-5.
29. Stram DO, Lee JW. Variance components testing in the longitudinal mixed effects model. *Biometrics* 1994; **50**(4):1171–1177. DOI: 10.2307/2533455.
30. Pinheiro J, Bates D. *Mixed-effects Models in S and S-Plus*. Springer: Berlin, 2000.
31. Wienke A, Arbeev K, Locatelli I, Yashin A. A comparison of different bivariate correlated frailty models and estimation strategies. *Mathematical Biosciences* 2005; **198**:1–13. DOI: 10.1016/j.mbs.2004.11.010.
32. Elbers C, Ridder G. True and spurious duration dependence: the identifiability of the proportional hazard model. *Review of Economic Studies* 1982; **49**:403–409. DOI: 10.1016/0014-2921(83)90056-9.
33. Diekmann O, Heesterbeek J, Metz J. On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations. *Journal of Mathematical Biology* 1990; **28**:365–382. DOI: 10.1007/BF00178324.

Copyright © 2009 John Wiley & Sons, Ltd.

34. Goethals K, Janssen P, Duchateau L. Frailty models and copulas: similarities and differences. *Journal of Applied Statistics* 2008; **35**(9):1071–1079. DOI: 10.1080/02664760802271389.
35. Wang W, Ding A. On assessing the association for bivariate current status data. *Biometrika* 2000; **87**:879–893. DOI: 10.1093/biomet/87.4.879.
36. White H. Maximum likelihood estimation of misspecified models. *Econometrica* 1982; **50**:1–25.
37. Litière S, Alonso A, Molenberghs G. Type I and Type II error random-effects misspecification in generalized linear mixed models. *Biometrics* 2007; **63**:1038–1044. DOI: 10.1111/j.1541-0420.2007.00782.x.
38. Litière S, Alonso A, Molenberghs G, Geys H. The impact of a misspecified random-effects distribution on the estimation and performance of inferential procedures in generalized linear mixed models. *Statistics in Medicine* 2008; **27**:3125–3144. DOI: 10.1002/sim.3157.

Copyright © 2009 John Wiley & Sons, Ltd.